Standard Setting Lessons Learned in the South African Context: Implications for International Implementation

Mary J. Pitoniak
Educational Testing Service

Nan Yeld
Centre for Higher Education Development, University of Cape Town ,South Africa

Abstract
Criterion-referenced assessments have become more common around the world, with performance standards being set to differentiate different levels of student performance. However, use of standard setting methods developed in the United States may be complicated by factors related to the political and educational contexts within another country. In this article, experience gained from conducting several standard setting studies in South Africa is shared. The legacy of the apartheid era, in which segregation and discrimination were institutionalized, affects the attitudes of South Africans toward assessment and placing students into performance levels. These issues played out as panelists were asked to make judgments related to students' likely performance in higher education. Although the instantiation of panelists' reluctance to label students may be different in South Africa compared to the United States or other countries, lessons can be learned about how the effects of these beliefs and anxieties may be addressed during standard setting activities.

Keywords: South Africa, standard setting, validity

Standards-based education plays a key role in many countries around the world. Comparisons are no longer limited to those that rank students; instead, students are evaluated by the degree to which they have learned the content required for promotion, graduation, or admission. One important component of such criterion-related

assessments is that of setting performance standards. Students' performance is compared to a set of knowledge, skills, and abilities seen to be needed to attain a given standard, and judgmental studies are then conducted to determine the cut scores, which are the numerical points on the score scale that will separate students into different performance categories.

Standard setting theory and research have received much attention in the United States over the past 30 years. However, it is less clear how the extant standard setting methods can be adapted for use in international settings. Will the methods work as intended? What aspects of the procedures may need to be modified when implementing standard setting methods in different political and educational contexts?

In this article, we describe pilot and operational standard setting studies conducted with a testing program in South Africa. Although that country—as every country—has unique issues, we believe that lessons learned in the South African context during these studies can be useful to others considering implementing judgmental standard setting studies in a non-American setting.

South African political and educational context

Every large-scale assessment takes place within a political and educational context. In the United States, for example, the No Child Left Behind Act (U.S. Department of Education, 2002) set forth educational goals for states to meet that are measured in part through the use of assessments. To provide the context for the issues discussed in this article, a brief description of the political and educational factors in South Africa will be provided.

Legacy of Apartheid

Apartheid was imposed in South Africa from 1948 to 1994. Racial segregation and political and economic discrimination against non-European groups were official policies of the government. An infamous quote in 1953 from the Minister of Native Affairs, Dr. Hendrik Verwoed, illustrates the views prevalent at that time:

> There is no place for [the Bantu] in the European community above the level of certain forms of labour. What is the use of teaching the Bantu child mathematics when it (sic) cannot use it in practice? That is quite absurd. Education must train people in accordance with their opportunities in life, according to the sphere in which they live. (Verwoerd, 1960, as quoted by Ratshitanga, 2007, p. 15)

The long-lasting and seemingly irreparable damage done by the repressive policies of the apartheid era is still very visible in South Africa in the educational sphere, at all levels. The most recent statistics available on performance of South

African secondary school students on the Trends in International Mathematics and Science Study (TIMSS) global assessments in mathematics (Mullis, Martin, Gonzalez, & Chrostowski, 2004) and science (Martin, Mullis, Gonzalez, & Chrostowski, 2004) confirm the lasting impact of the differential funding approaches of the past, where, at the height of apartheid, state spending on Black schools was one tenth of that spent on White school education (Byrnes, 1996). For example, for students at schools attended almost entirely by African (Black) students, the average scores on TIMSS were 277 for mathematics and 199 for science (Reddy, 2003). For students at historically, and still predominantly, White schools, the corresponding scores were 468 and 483. In other words, White school achievement was on a par with the international average of 467 for mathematics and 474 for science, but that of students at Black schools was less than half this level.

A cursory investigation of infrastructural provision at schools sheds some light on why performance remains so seriously skewed. In a national survey conducted in 2011 (Department of Basic Education, 2011a), it was found that of 24,793 publicly-funded (state) schools, 3544 had no electricity, 2402 no water, 913 no ablution facilities, 19,541 no libraries, and 21,021 no laboratories. Since the overwhelming majority of the schools without such facilities are those formerly designated as schools for Black children, it can be seen that the continuing inequities still impact more severely on opportunities and provision for those targeted by apartheid.

While unsurprising, the close association of income and race and its impact on achievement bears comment. For example, performance on the public National Senior Certificate (the national school-leaving examination) varies by the degree of poverty in the secondary schools that the students attend. In South Africa, state schools are categorized into quintiles based on rates of income, unemployment, and illiteracy within the school catchment area. One statistic of note here is the percentage of schools in which 80% or more of the students obtained the school-leaving certificate. In the poorest quintile of schools only 18% of the schools attained this benchmark; in the wealthiest quintile, 64% of the schools did (van der Bergh, 2008). One of the major issues in relation to performance is that the poorer schools are almost 100% Black, while the wealthiest schools are predominantly White. Poverty and race are thus closely associated, and the effects play themselves out in academic achievement so that failure becomes, in effect, a Black phenomenon. The consequences of this are powerful and varied and impact on the attitudes brought to the standard-setting task.

Performance statistics in the higher education sector show the persistence of this educational disadvantage. For example, the “Student Pathways” study undertaken by Letseka, Cosser, Breier, and Visser (2010) reported that about 40% of students at South African public institutions drop out of their studies during or at the end of their first year, and that only about 15% of students admitted to higher education obtain their degrees in the minimum time. Both this study and others

such as that conducted by Scott, Yeld, and Hendry (2007) are based on cohort studies, tracking students from entry to completion or exit, and both show striking and disturbing differences in graduation rates between White and Black students. Basing their findings on students enrolling in contact, residential universities, Scott and colleagues estimated that the most optimistic scenario for graduation was about 44%, and very much lower when distance education was included.

This low graduation rate is not unique in the world. What makes the South African situation so striking is that the students that make it into higher education are a highly selected elite, representing the survivors of an extremely poor K-12 schooling system. The severity of the attrition in schooling is illustrated in Figure 1, which shows that only 5% of the students who entered schooling obtained results that made them eligible to enter higher education for degree study purposes (a further 16% were eligible for diploma or certificate study). That less than half of this very small percentage—about 2.5%—obtained a Bachelor's degree (Scott et al., 2007) shows the severity of the problems facing higher education in South Africa.

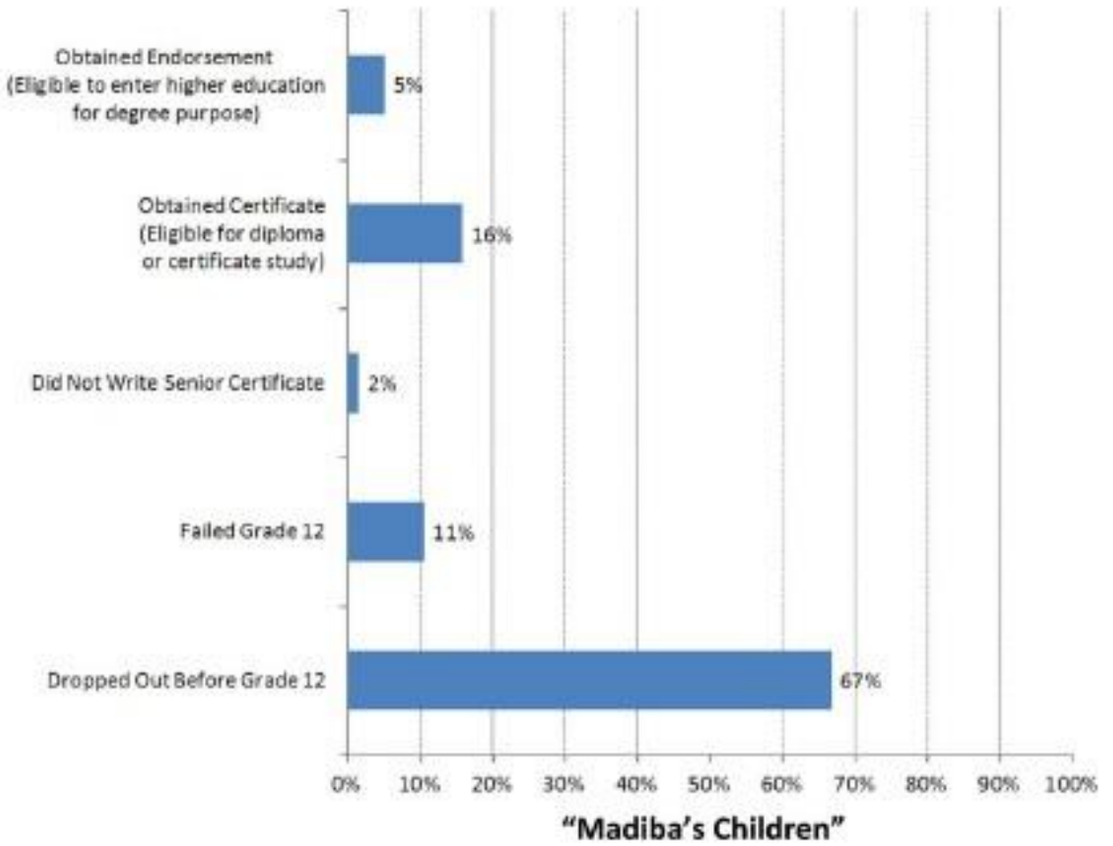

Figure 1

Educational outcomes of South African students who entered school in 1995; the lowest education outcome is at the bottom of the graph and the highest is at the top. Madiba is the clan name of former President Nelson Mandela; it is used in this graph as it deals with the students who entered schooling the year after Mandela became president (color figure available online).

Attitudes Toward and Understanding of Assessment

There are pervasive and lingering suspicions about assessment in South Africa. The society is highly sensitive to judgments, which are seen as a means of division and discrimination. While this is of course true in that all assessment aims to make judgments about performance, the country's past was dominated by bias and prejudice, and this has engendered lasting suspicion.

There is also less awareness of criterion-referenced assessment in South Africa. In general, teachers within South Africa may be more familiar with norm-referenced interpretations than criterion-referenced ones. The "big test" in South Africa, the National Senior Certificate, is itself very much a norm-referenced set of examinations, with its assessment results routinely standardized to conform to acceptable distributions. This tends to mask the extent of the educational challenge in the country (Yeld, 2011), since accurate criterion-referenced data about the performance of students are not made available.

There is no tradition within the country of standards-based assessments that place students into achievement levels such as those reported for the U.S. National Assessment of Education Progress (see, e.g., National Center for Education Statistics, 2011). However, the South African Department of Basic Education launched in 2011 the first fully-fledged application of the Annual National Assessments program, a set of criterion-referenced tests administered at key stages of schooling in mathematics and languages. This launch followed trial runs in 2008 and 2009, necessitated in large part by the need to build capacity in the system to understand standardized assessment approaches (Department of Basic Education, 2011b). This focus on building assessment expertise among teachers and administrators might well mean that panelists in future standard-setting studies have greater understanding of standards-based approaches.

In addition, the lack of awareness of the results or process of standard setting is a factor in the country. While U.S. standard setting panelists are likely more familiar with the classification of students into performance categories and may even know someone who has participated in a standard-setting study, this was not the case in South Africa. Thus the entire context of the use of the assessment, and the purpose and procedures used to develop the benchmarks, or cut scores, may be totally novel to the panelists, as was the case in this study.

The National Benchmark Tests and the Standard-setting Process

Within this section, the test for which standards were set will be described. A brief summary of the standard setting process used is also provided.

The National Benchmark Tests

In 2005, Higher Education South Africa, the organization representing all higher education institutions in the country, resolved to develop several strategies to address the very high levels of failure. One of these strategies was to develop a set of assessments in core domain areas that would accurately identify the educational needs of the sector's incoming students (Higher Education South Africa, 2006). Armed with this knowledge, it was assumed, faculty would be able to design more appropriate curricula and also to design curricular routes that would incorporate remedial and developmental elements, and as a consequence, academic progression and graduation rates would improve.

The Centre for Higher Education Development at the University of Cape Town was commissioned to develop the assessments. The project became known as the National Benchmark Tests Project (NBTP). Following wide consultation and sector participation in all processes, the NBTP became fully operational in 2009, testing students' proficiencies in three core areas believed to underlie future academic success: Academic Literacy, Quantitative Literacy, and Mathematics (see Table 1 for a brief description of these domains).

The performance levels for the NBTs are presented in Table 2. Students at the Proficient level would be able to complete a regular program of study. Those at the Intermediate level would face challenges, and may need remediation or an extended program of study. Students at the Basic level would face serious learning challenges and would need extensive and long-term support, perhaps provided by a bridging program, to attend university.

Standard Setting Studies

Through a collaborative arrangement with Educational Testing Service (ETS), the NBTP worked with ETS in the standard-setting process. A pilot standard-setting

Table 1. Description of National Benchmark Test Domains

Test	Brief Description of Domain
Academic Literacy	The extent to which students can cope with typical reading and writing demands in English (at this stage)
Quantitative Literacy	The ability to manage situations or solve problems of a quantitative nature in practice, and to respond to quantitative information represented in various modes
Mathematics	The ability to manipulate, raise questions, synthesize a number of different mathematics concepts, and draw strictly logical conclusions in abstract symbolic and complex contexts

Table 2. Generic Performance Level Descriptions for National Benchmark Tests

Level	Description
Proficient	Performance in domain areas suggests that subsequent academic performance will not be adversely affected. If admitted, students should be placed on regular programme of study.
Intermediate	Challenges in domain areas identified; it is predicted that academic progress will be affected. If admitted, students' educational needs should be met in a way deemed appropriate by the institution (e.g., extended or augmented programmes, special skills provision).
Basic	Serious learning challenges identified; it is predicted that students will not cope with degree level study without extensive and long-term support, perhaps best provided through bridging programmes or FET [Further Education and Training]. Institutions registering students performing at this level would need to provide such support.

study was conducted in 2008, and an operational study in 2009. The Angoff (1971) standard setting method was used. After training was provided, which included panelists' taking the test, two rounds of ratings were conducted. After round 1, feedback was given to the panelists, and discussions took place. Feedback included rater location feedback (information on how panelists' ratings compared to each other, at both the item and total test score level), performance data (p-values), and impact data (percentage of students placed into each of the achievement levels).

The number of panelists per panel for the operational standard setting study was 21 for Academic Literacy, 15 for Quantitative Literacy, and 14 for Mathematics. As noted by Hambleton and Pitoniak (2006), demographic targets should be set for panel composition, and the panels should contain representatives from each of the groups that will be affected by the outcome of the assessments and the decisions that will be made based on the results. Several targets were set for these standard setting studies.

In South Africa, ethnicity is an important consideration given the history of the country, and efforts were made to panelists representing the four officially designated population groups: White, Black, Coloured,[1] and Indian.[2] In Table 3,

[1] The term Coloured is used in South Africa to refer to an ethnic group of mixed-race people who possess some African ancestry but not enough to be considered Black under the law of South Africa. The use of "race-based" categories is strictly guided by the need to provide redress for historical discrimination, and for equity employment purposes.

[2] The term Indian is used to refer to people of Indian descent living in South Africa, with India referring to the South Asian country.

Table 3. Number of Panelists by Ethnicity

Ethnicity	Academic Literacy	Quantitative Literacy	Mathematics
		Pilot Study	
Black	15% (2)	9% (1)	9% (1)
Coloured	16% (2)	9% (1)	0% (0)
Indian	0% (0)	9% (1)	0% (0)
White	70% (9)	73% (8)	91% (10)
		Operational Study	
Black	14% (3)	0% (0)	7% (1)
Coloured	10% (2)	7% (1)	7% (1)
Indian	10% (2)	7% (1)	7% (1)
White	67% (14)	87% (13)	80% (11)

information about the ethnicity of panelists from both the pilot and operational standard setting studies is presented. The majority of the panelists were White, reflecting the racial composition of university faculty. However, all but one of the six panels (three each for pilot and operational) had at least one Coloured and one Black panelist, and all but two panels had at least one Indian panelist.

It was also seen as critical that panelists be representative of different geographic regions and types of universities, with type of university often being a proxy for the primary racial group composing the student body (during apartheid, universities were created to cater to the needs of different linguistic and ethnic groups). Representation from a wide range of subject areas was also a factor, since the nature of the domains meant that a wide range of disciplines were included for each test. Gender was also taken into account.

Since the focus of this article is on challenges raised by the standard setting endeavor, not the outcome of the studies, additional information about the methodology and results will not be presented. Further information about the standard setting studies can be found in Pitoniak and Yeld (2011).

Challenges faced in the Standard Setting Process

The factors described previously—including the political and educational context, attitudes toward assessment, and the purpose of the NBPT—undeniably played a role in the standard setting process. Issues that arose during the standard-setting studies are discussed in this section.

Placing Students into Performance Levels

The outcome of a standard setting session is recommended cut scores that will separate students into different performance levels. Even conceptualizing this task was problematic for some panelists, for reasons stemming from the political and educational contexts described previously.

The reality that the poorest schools (and universities) are almost entirely attended by Black students means that the great majority of poor performers on the NBTP—for example, those placed into the Basic category, deemed unlikely to succeed at university even with support—will be Black. This was difficult for panelists to accept. The enduring injustice of unequal opportunities to learn has made educators feel that failing students (or categorizing them as Basic or Intermediate) is simply a case of blaming the victims and further disadvantaging them. There are fears that "educational underpreparedness" will be conflated with "lack of intelligence" and further confirm the stereotype, so assiduously cultivated by apartheid, that Black people are less capable of higher order thinking than others.

Panelists also voiced fear that labeling performance as Basic, for example, will lead to students in this category being excluded or not admitted to higher education, with its potential for greatly enhanced social and economic upward mobility. Such anxiety appeared to stem largely from reluctance to jeopardize a student's chance of being admitted to a university. This was very difficult to address since the reality is that most institutions in fact do not provide the support that is needed, and thus might be tempted to take the easy way out and not admit the candidate with a Basic score.

Panelists were reminded that the aim of the NBTP is precisely to try to find out which students need support, and how much support. They were encouraged to overcome their reluctance to provide judgments about students, and were reminded that unless the problems and needs are made visible, institutions can continue to ignore them, and high failure rates will persist. It was also pointed out to panelists that South African participation rates, particularly those of Black students, are very low and there is pressure to increase them. Therefore, pragmatically, institutions are very unlikely to be able, even if they wished to do so, to turn away otherwise qualified candidates who score in the Basic category.

Panelists also had a lot of questions about the meaning of the performance levels, both in general and specific to their subject areas. General concerns related to the definitions of program completion on a normal schedule, and the nature of remediation. For example, for the Proficient level panelists had to be reminded that students in that level would not necessarily excel at university, but would instead be able to complete a regular program of study. For the Intermediate level, panelists voiced concern that the remediation needed by students at that level would not in fact be offered, and were asked to make their judgments assuming that it would

be. Some panelists had a lot of difficulty making this assumption. The following panelist comment is relevant to this issue.

> To minimize anxiety, I think we should be made aware early on that the Department of Education has funding earmarked for support because the concern for me was in between the two benchmarks: If support is not provided, I'd rather put more people into basic (i.e., raise benchmarks) so my initial ratings were perhaps inflated as I don't know if support is available.

The difficulty that panelists had with the very goal of the study was made apparent not only through discussions when doing the rating task, but during most parts of the process. Some of those effects are discussed in the following sections.

Retaining Information about the Task

Throughout the pilot study, particularly on the first day of the session, panelists repeatedly asked questions to which the answers had already been provided. It quickly became clear that the timing and procedures that the facilitator had adopted from studies conducted in the United States could not be transferred wholesale to South Africa.

Language was likely not the cause since the panelists spoke English, sometimes in addition to another of the 11 official and numerous unofficial South African languages. One possible reason was that the information was not presented clearly or often enough. However, it may also have been the case that panelists' anxiety levels were so high that they were not retaining the information.

Most standard setting facilitators have had the experience of providing training, only to have a panelist ask when one of the very topics just presented will be discussed. In South Africa, this experience was much more common. Therefore, in the operational studies, more time was allotted for training, and concepts were repeated far more often. Multiple handouts that the panelists could keep at their desks were also provided. As in any standard setting study, feedback from questionnaires provided to panelists throughout the process was used to provide remediation prior to the start of the next activity.

Providing Round 2 Ratings

As noted previously, there were two rounds of ratings in the study. The feedback presented after round 1 consisted of rater location feedback (information on how panelists' ratings compared to each other, at both the item and total test score level), performance data (p-values), and impact data (percentage of students placed into each of the achievement levels). In the pilot study, the facilitator instructed

panelists to make round 2 ratings taking these sources of information into account, but panelist behavior was not quite as expected.

Some panelists proceeded to take only several minutes in which they barely glanced at the information, revised only one or two ratings at most, and then closed their test booklets, stopping all activity. The statement that they were not required to revise their ratings was taken as allowance to not look at the items or review ratings at all. Again, this behavior could have been due in part to anxiety over the rating process or a lack of understanding of same. The instructions were made much firmer in the operational study. It was stressed to panelists that they were required to review each item during the second round of ratings and make a conscious decision whether to revise their estimates of borderline performance, and their behavior was monitored to see if they were in fact performing the task. This increased the level of procedural compliance.

Lessons Learned

Although the standard setting study design incorporated the common steps called for in such activities (Hambleton, Pitoniak, & Copella, 2011; Pitoniak & Morgan, 2011), implementation in South Africa was complicated by the issues described previously. Heeding the following lessons learned may reduce the impact of these factors in standard setting conducted under these conditions.

- Prior to the standard setting session, the sponsoring agency and the facilitator(s) should discuss the political and educational implications of the test and its results so that preparations can be made for their possible impact on the standard setting study.
- A pilot standard setting study should be conducted to assess how well the procedures for a given method can be implemented in a given context, and how they may need to be modified for operational use.
- The overview provided to panelists must make clear the purpose and possible outcomes of the assessment, particularly for a new testing program. Issues related to the possible impact of the results on different constituencies, if known, should be acknowledged.
- Panelists should be given ample time to voice their opinions about the testing program, while being reminded that the policies have been set and their role is to perform the tasks called for in the standard setting study.
- The overview and training should directly acknowledge feelings that may arise when panelists are asked to place students into performance categories.
- Extended time should be allowed for training in contexts in which the placement of students into performance levels is more politically charged.
- Additional time should be provided for all other tasks, if possible.

- Because panelists may be challenged by keeping key concepts in mind while performing the rating task, more handouts should be prepared for content that panelists may need to reference during the study.
- Panelists should be given explicit reminders of the steps to be followed during each task, even if they have been outlined thoroughly in previous steps.

Conclusions

The overall lesson that was learned, or rather reinforced, in the South African standard setting studies is that the educational context of a given country has a large impact on all aspects of large-scale assessment. Within South Africa, the educational system reflects the changing social and cultural conditions of the postapartheid years. Inequities are being addressed, but there are challenges and realities that must be faced. It needed to be acknowledged during the studies that making judgments about student performance may result in uncomfortable feelings and differences of opinion, but that the standard setting task is one that will ultimately support decisions that will be made in the students' best interests.

While issues unique to the South African context amplified difficulties during various steps of the process, the results can provide useful reminders to anyone conducting a standard setting study. Panelists need to be given very clear information about the purpose of the standards, sufficient time must be provided so that they will absorb information, and the steps should be structured such that their understanding of the task is continually reinforced. Heeding these reminders will enhance the validity of the interpretations made on the basis of the classifications into the performance levels.

Every part of the standard setting endeavor—identification of stakeholders, recruitment of panelists, choice of method, procedural steps for a given method, etc.—is affected by the climate in which the assessment is given and the results will be reported. Consideration of related issues must be undertaken as the study is planned, and before specific design decisions are made.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–597). Washington, DC: American Council on Education.

Byrnes, R. M. (1996). South Africa: A country study. Washington, DC: GPO for the Library of Congress.

Department of Basic Education. (2011a). NEIMS (National Education Infrastructure Management System) report 2011. Pretoria, South Africa: Author.

Department of Basic Education. (2011b). Report on the annual national assessments of 2011. Pretoria, South Africa: Author.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. (2011). Setting performance standards. In G. Cizek (Ed.), Setting performance standards: Theory and applications (2nd ed., pp. 47–76). New York, NY: Routledge.

Higher Education South Africa. (2006, May). Access and entry level benchmarks: The national benchmark tests project. Pretoria, South Africa: Author.

Letseka, M., Cosser, M., Breier, M., & Visser, M. (2010). Student retention and graduate destination: Higher education and labour market access and success. Pretoria, South Africa: Human Sciences Research Council.

Martin, M. O., Mullis, I. V S., Gonzalez, E. J., & Chrostowski, S. J. (2004). TIMSS 2003 international. science report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/timss2003i/scienceD.html

Mullis, I. V S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). TIMSS 2003 international. mathematics report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/timss2003i/mathD.html

National Center for Education Statistics. (2011). The Nation's Report Card: Reading 2011 (NCES 2012–457). Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012457.pdf

Pitoniak, M. J., & Morgan, D. L. (2011). Setting and validating cut scores for tests. In C. Secolsky & D. B. Denison (Eds.), Handbook on measurement, assessment, and evaluation in higher education (pp. 343–381). New York, NY: Routledge.

Pitoniak, M. J., & Yeld, N. (2011, April). Standard setting lessons learning in the South African context. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Ratshitanga, M. (2007). Our present is indeed connected to our past. Pretoria News, March 20: 15.

Reddy, V (2003). Mathematics and science achievement at South African schools in TIMSS in 2003.. Pretoria, South Africa: Human Sciences Research Council Press.

Scott, I. R., Yeld, N., & Hendry, J. (2007). Higher Education Monitor 6: A case for improving teaching and learning in South African higher education. Brummeria, South Africa: Council on Higher Education.

U.S. Department of Education. (2002). No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425.

van der Bergh, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. Studies in Educational Evaluation, 34(3), 145–154.

Yeld, N. (2011, January). Confusion, suspicion continue to plague our matric results. Business Day, January 12.